

Minimal Perceptrons for Memorizing Complex Patterns

Marissa Pastor^{a,b,1}, Juyong Song^{a,c,1}, Danh-Tai Hoang^{a,d}, Junghyo Jo^{a,c,*}

^a*Asia Pacific Center for Theoretical Physics, Pohang, Korea*

^b*Department of Physics, University of San Carlos Talamban, Cebu, Philippines*

^c*Department of Physics, Pohang University of Science and Technology, Pohang, Korea*

^d*Department of Science, Quang Binh University, Dong Hoi, Vietnam*

Abstract

Feedforward neural networks have been investigated to understand learning and memory, as well as applied to numerous practical problems in pattern classification. It is a rule of thumb that more complex tasks require larger networks. However, the design of optimal network architectures for specific tasks is still an unsolved fundamental problem. In this study, we consider three-layered neural networks for memorizing binary patterns. We developed a new complexity measure of binary patterns, and estimated the minimal network size for memorizing them as a function of their complexity. We formulated the minimal network size for regular, random, and complex patterns. In particular, the minimal size for complex patterns, which are neither ordered nor disordered, was predicted by measuring their Hamming distances from known ordered patterns. Our predictions agreed with simulations based on the back-propagation algorithm.

Keywords: perceptrons, network complexity, binary patterns, memory storage, network architecture

1. Introduction

Neural signaling in synaptic networks motivated the early study of artificial neural networks, to recapitulate the learning capability of the brain.

*Corresponding author

Email address: jojunghyo@apctp.org (Junghyo Jo)

¹These authors contributed equally to this work

Their utility has expanded from that inception to serving as alternatives to conventional computers for input/output processing or as exemplars of parallel distributed processing, and they have been successfully applied to pattern classification [1, 2] and memory storage [3, 4, 5, 6]. Standard implementations of neural networks map inputs \mathbf{x}^μ to outputs z^μ through intermediate processing layers. The M input/output pairs, $\xi^\mu = (\mathbf{x}^\mu, z^\mu)$, form a *pattern*, $\boldsymbol{\xi} = \{\xi^1, \xi^2, \dots, \xi^M\}$. This input/output mapping can be achieved in two different ways: The neural network can either (i) learn the underlying rule for the mapping from some training pairs, or (ii) memorize the whole pattern of input/output pairs and retrieve the stored outputs in response to given inputs. In either way, it is a major impediment that the required complexity of network architectures for learning or memorizing certain patterns is, in general, unknown. In this paper, we focus on memorizing patterns.

Designing the optimal network architecture has been known as an NP (Non-deterministic Polynomial-time) problem that requires computationally expensive search techniques and optimization [7, 8, 9, 10]. Indeed, most attempts use empirical approaches and proceed by scanning over different network configurations while utilizing incremental [11] and/or pruning algorithms [12, 13]. Simple networks may lead to insufficient memory and poor generalization, while complex networks lead to poor predictive performance by overestimating each element in patterns [6]. The required complexity of networks generally depends on the complexity of patterns for memorizing. Therefore, if the complexity of patterns and networks could be quantified, the optimal network architecture could be systematically designed.

Two popular complexity measures for patterns are Shannon entropy, the degree of uncertainty for describing a pattern [14, 15, 16], and Kolmogorov complexity, the length of the shortest computer program for generating a pattern [17]. However, the uncertainty or probability of each element in a pattern is unknown and the algorithmic complexity is itself difficult to compute. These difficulties suggest the need for a metric to quantify the practical complexity of patterns relevant for perceptrons. Here we propose a simple complexity index and relate it to the minimal network size for memorizing patterns.

This paper is organized as follows: We introduce the mathematical description of our feedforward neural network in Sec. 2, and a storage problem of binary patterns of different complexities in Sec. 3. Next, we estimate minimal network sizes for storing regular binary patterns in Sec. 4.1 and random binary patterns in Sec. 4.2, and compare them to simulation results. Then,

we generalize the complexity formulation to estimate minimal network size for storing complex binary patterns in Sec. 5. Finally, we summarize the paper in Sec. 6.

2. Neural network

We study a three-layer feedforward neural network as shown in Fig. 1. This simple network architecture is successful at solving pattern recognition problems [18, 19]. In addition, the universal approximation theorem proves that the three-layer network suffices to approximate any continuous function, $z^\mu = f(\mathbf{x}^\mu)$ [20]. For simplicity, we consider N -dimensional vectors of binary inputs $\mathbf{x}^\mu = (x_1^\mu, x_2^\mu, \dots, x_N^\mu)$ and scalar binary outputs z^μ . One pattern ξ represents 2^N pairs of (\mathbf{x}^μ, z^μ) , because each component in the N -dimensional input vector takes values $x_i^\mu = 0$ or 1 . The input/output mapping requires N input nodes and a single output node. In the feedforward three-layer network, an input \mathbf{x}^μ is transformed into the activities $\mathbf{y}^\mu = \{y_1^\mu, y_2^\mu, \dots, y_H^\mu\}$ of H hidden nodes:

$$y_j^\mu = \sigma\left(\sum_{i=1}^N w_{ji}x_i^\mu - w_{j0}\right), \quad (1)$$

where w_{ji} is the connection weight from the i th input node to the j th hidden node, and w_{j0} is the bias of the j th hidden node. With these definitions, the j th hidden node is activated when the integrated input signal $\sum_i w_{ji}x_i^\mu$ exceeds the bias w_{j0} through the sigmoidal activation function, $\sigma(a) = 1/(1 + e^{-a})$. The transformation from the hidden layer to the output layer follows the same rule:

$$z^\mu = \sigma\left(\sum_{j=1}^H v_j y_j^\mu - v_0\right), \quad (2)$$

where v_j is the connection weight from the j th hidden node to the output node, and v_0 is the bias of the output node. Successful storage of the pattern ξ represents the correct input/output transformation through the feedforward network with appropriate parameters (w_{ji}, v_j) , $i \in \{0, 1, \dots, N\}$ and $j \in \{0, 1, \dots, H\}$. The required minimum network size (i.e., number of hidden nodes, H) for the successful storage of a certain pattern is the key question we address.

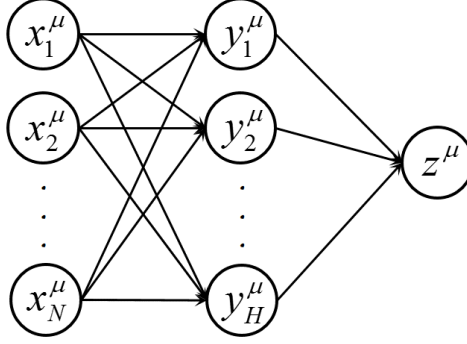


Figure 1: Three-layer feedforward neural network.

3. Pattern complexity

An example of the problem of binary pattern storage is a dichotomy problem on a binary N -cube. For a linearly separable problem, the transformation from inputs to the j th hidden node corresponds to the dichotomy $\{Y_j^+, Y_j^-\}$ of the elements \mathbf{x}^μ above and below an $(N - 1)$ -dimensional hyperplane, $\sum_{i=1}^N w_{ji}x_i^\mu - w_{j0} = 0$. Thus, for a simple binary pattern, one hyperplane is sufficient to dichotomize ξ^μ with respect to $z^\mu = 0$ and $z^\mu = 1$ elements (Fig. 2a). This means that one hidden node is sufficient to store the simple pattern. For more complex patterns (Fig. 2b), however, additional hidden nodes (or hyperplanes) and processing from the hidden layer to the output layer are necessary. The complexity of binary patterns is lower as neighboring elements are homogeneous in the binary N -cube. This observation suggests a simple complexity index:

$$K = \frac{1}{M} \sum_{\mu=1}^M k^\mu, \quad (3)$$

where k^μ is the number of different neighboring elements for an element ξ^μ . Considering the binary vectors $\mathbf{x}^{\mu'}$, one-bit different from the vector $\mathbf{x}^\mu = (x_1^\mu, x_2^\mu, \dots, x_N^\mu)$, as the neighborhood set Λ_μ of \mathbf{x}^μ , the individual complexity index is defined as $k^\mu = \sum_{\mu' \in \Lambda_\mu} (1 - \delta_{z^\mu, z^{\mu'}})$, using the Kronecker delta function $\delta_{z, z'} = 1$ for $z = z'$, and 0 otherwise.

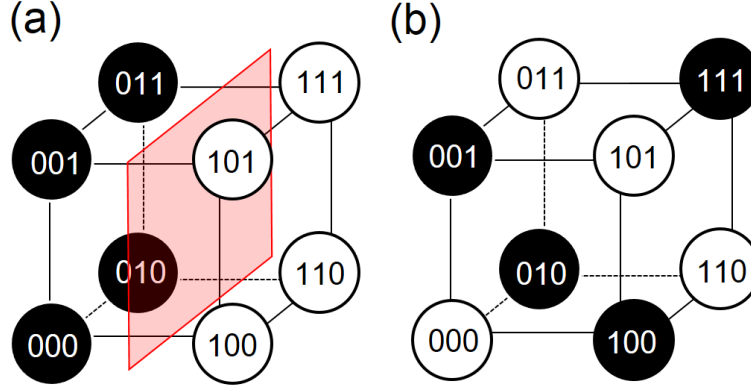


Figure 2: (Color online) Binary patterns. (a) A simple pattern and (b) a complex pattern for an input dimension $N = 3$. A hyperplane (red) separates two groups of black and white elements.

4. Minimal network for regular and random patterns

4.1. Regular patterns

The minimum number of hidden nodes required for memorizing some regular patterns is known. Parity patterns, $z^\mu = \text{mod}(\sum_{i=1}^N x_i^\mu, 2)$, require $H = N/2 + 1$ and $H = (N + 1)/2$ hidden nodes for even and odd N input dimensions, respectively [21, 22]. Parity patterns have the complexity index $K = N$, because every neighbor of an element has different outputs despite one-bit difference in their inputs. One may generate simpler regular patterns by introducing pseudo bits that have no effect on the output. When n pseudo-bits are introduced amongst the input bits such that the effective inputs become $(N - n)$ bits, this (N, n) pseudo-parity pattern requires fewer hidden nodes for memorizing the pattern. The pseudo pattern has a reduced complexity index $K = N - n$. For such regular patterns, the minimum number of hidden nodes is generally formulated as

$$H = \frac{K}{2} + 1 \quad (4)$$

for even N and $H = (K + 1)/2$ for odd N .

4.2. Random patterns

The complexity index K measures local complexities in patterns. Thus it is insufficient if we want to capture global order underlying patterns. Parity patterns look less complex than random patterns, although they have the highest complexity index. Such patterns have a strong order or rule that makes numerous elements, having identical values for $\sum_{i=1}^N (-1)^i x_i^\mu$, redundant [21]. Random patterns, lacking any order, do not have redundant elements for memorizing. Therefore, it is to be expected that random patterns require the maximum number of hidden nodes for memorizing, given K . Although the storage of random patterns itself may be practically useless, it determines the upper bound of H for the storage of binary patterns.

Here we propose one strategy for memorizing random patterns. Suppose that a random binary pattern has a total of $M (= 2^N)$ elements in which pM elements are black ($z^\mu = 1$), and $(1-p)M$ elements are white ($z^\mu = 0$). First, we introduce a reference hyperplane Y_1 dividing the N -cube into two regions, Y_1^+ and Y_1^- above and below the hyperplane, that contain pM and $(1-p)M$ elements, respectively (Fig. 3). Then, the Y_1^+ region has $(1-p)pM$ impurities of white elements, while the Y_1^- region has $p(1-p)M$ impurities of black elements.

Next, by introducing a pair of hyperplanes, Y_2 and Y_3 , we can isolate clustered impurities above/below the reference hyperplane. As shown in Table 1, the paired hyperplanes can selectively correct the outputs of the impurities without perturbing the outputs of the other elements. In principle, we can define the paired hyperplanes from clustered impurities. Suppose we choose $(N+1)$ impurities ξ^μ to define a Y_2 hyperplane. These impurities imply $(N+1)$ linear equations, $\sum_{i=1}^N w_{2i} x_i^\mu - w_{20} = 0$, which can also be described as a matrix equation, $\mathbf{X}_2 \cdot \mathbf{w}_2 = \mathbf{0}$. Similarly we can define a Y_3 hyperplane with $\mathbf{X}_3 \cdot \mathbf{w}_3 = \mathbf{0}$ by choosing another $(N+1)$ impurities which are located sufficiently close to the Y_2 hyperplane. Then we can place those $2(N+1)$ impurities into a small interspace between the Y_2 and Y_3 hyperplanes by slightly rotating the two hyperplanes in opposite directions with respect to their intersection (Fig. 3): $\mathbf{X}_2 \cdot \mathbf{w}_2 = \boldsymbol{\epsilon}_2$ and $\mathbf{X}_3 \cdot \mathbf{w}_3 = \boldsymbol{\epsilon}_3$. Here we can uniquely determine the paired hyperplanes for isolating $2(N+1)$ impurities, if both matrices \mathbf{X}_2 and \mathbf{X}_3 have rank $(N+1)$. This implies that one additional hyperplane can isolate $(N+1)$ impurities. To isolate a total of $2p(1-p)M$ impurities, we need $2p(1-p)M/(N+1)$ hyperplanes. Therefore, the minimum number of hidden nodes for memorizing random patterns corresponds to the total number of the impurity hyperplanes in addition to the single reference

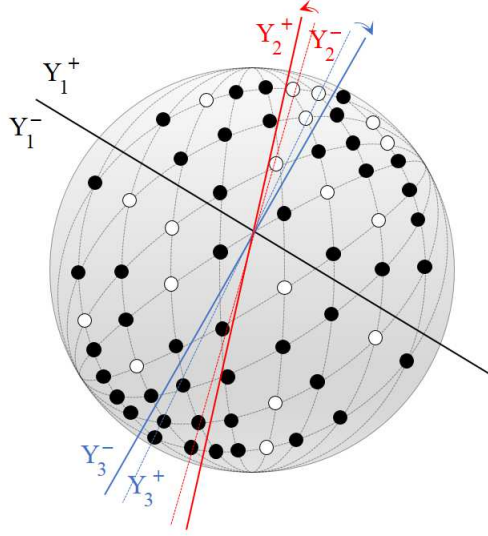


Figure 3: (Color online) Schematic diagram for the storage of binary random patterns. A reference hyperplane Y_1 is introduced for the dichotomy of black and white elements in Y_1^+ and Y_1^- , respectively. Additional hyperplanes Y_2 and Y_3 isolate impurities of white elements in Y_1^+ and black elements in Y_1^- .

hyperplane:

$$H = \frac{2^N K}{N(N+1)} + 1, \quad (5)$$

where the complexity index $K = 2p(1-p)N$ is derived from Eq. (3). The random pattern has pM black elements with $k^\mu = (1-p)N$, and $(1-p)M$ white elements with $k^\mu = pN$ on average.

The minimum number H for random patterns can be further reduced if we use unexpected relations between the conjugate elements ξ^μ and $\xi^{\mu*}$ linked by their inputs as $\mathbf{x}^\mu + \mathbf{x}^{\mu*} = (1, 1, \dots, 1)$. Suppose that we choose N impurity elements for which N conjugate elements have opposite colors ($z^\mu \neq z^{\mu*}$). If the conjugate elements are also impurities located in the opposite side of the reference hyperplane Y_1 , one hyperplane is sufficient to define $2N$ impurity elements. These conjugate impurity elements have $2N$ equations: $\sum_{i=1}^N w_{ji} x_i^\mu - w_{j0} = 0$ and $\sum_{i=1}^N w_{ji} x_i^{\mu*} - w_{j0} = 0$. However, the conjugate symmetry reduces $2N$ equations to $(N+1)$ equations: $\sum_{i=1}^N w_{ji} x_i^\mu - w_{j0} = 0$ and $\sum_{i=1}^N w_{ji} - 2w_{j0} = 0$. Thus, the $(N+1)$ variables w_{ji} can be uniquely

Table 1: Dichotomy of random binary patterns. The weight parameters are chosen as $v_j = 2$ and the bias as $v_0 = 3$.

	y_1	y_2	y_3	$\sum_j v_j y_j - v_0$
$Y_1^+ Y_2^+ Y_3^-$	1	1	0	+1
$Y_1^+ Y_2^- Y_3^+$	1	0	1	+1
$Y_1^+ Y_2^- Y_3^-$	1	0	0	-1
$Y_1^- Y_2^+ Y_3^-$	0	1	0	-1
$Y_1^- Y_2^- Y_3^+$	0	0	1	-1
$Y_1^- Y_2^+ Y_3^+$	0	1	1	+1

determined.

This is one scenario out of four possibilities depending on the color and location of the conjugate elements: (i) different color and opposite side to Y_1 ; (ii) same color and same side; (iii) different color and same side; and (iv) same color and different side. As explained for Case (i), Case (ii) can also define $2N$ impurities using one hyperplane. However, Cases (iii) and (iv) cannot use the conjugate symmetry, because the conjugate elements are not impurities in Y_1 . For these cases, we can define only $(N + 1)$ impurities with one hyperplane. Among $m(\equiv 2p(1 - p)M)$ impurities, $2p(1 - p)m$ impurities correspond to Cases (i) and (ii), while $[p^2 + (1 - p)^2]m$ impurities correspond to Cases (iii) and (iv). Thus the minimum number of hyperplanes is $H = 2p(1 - p)m/2N + [p^2 + (1 - p)^2]m/(N + 1) + 1$. This gives a second-order correction to Eq. (5):

$$H = \frac{2^N K}{N(N + 1)} \left[1 - \frac{(N - 1)K}{2N^2} \right] + 1. \quad (6)$$

The estimated H from Eqs. (4) and (6) for regular and random patterns, respectively, are tested by simulating supervised three-layer feedforward neural networks. Once a network with a certain number of hidden nodes is given, inputs \mathbf{x}^μ propagate to an output \tilde{z}^μ through the hidden layer, as described in Eqs. (1) and (2). Then we measure the mismatches between the computed outputs \tilde{z}^μ and the true outputs z^μ as an error, $E = 1/2 \sum_{\mu=1}^M (z^\mu - \tilde{z}^\mu)^2$. Here we optimize the weight and bias parameters with the back-propagation (BP) algorithm [23], a gradient-descent method that updates parameters according to the change needed for decreasing the error E . For example, the update of weights w_{ji} follows $\Delta w_{ji} = -\alpha \partial E / \partial w_{ji}$, in which the learning rate is optimized as $\alpha = 0.35$ for fine, but not slow, searches for the error-landscape

minima. We iterate the forward process and the error back-propagation by a few million times, and record the final error E after equilibrium is reached. Various size of hidden layers have been examined, and we have found the minimum H for a successful storage of patterns with an accuracy of $E < 2^{-N}$. We have considered more than 5000 ensembles starting from different initial parameter values, because the nonlinear feedforward equations have many local minima in the error landscape. Although BP algorithm is an old technique, its use in this study merits a direct relation between the network size and the complexity of the input signal patterns. Additionally, the extensive use of BP [6, 23, 18] and its various hybrid forms (modified BP or combination with other learning algorithms), imply that the minimal perceptrons to be found may give insights on the first principles of network size optimization for memory storage based on the complexity of input signals.

As shown in Fig. 4, the minimum H obtained in the simulation is consistent with the estimations of Eq. (4) for regular patterns and Eq. (6) for random patterns.

5. Minimal network for complex patterns

The strategies for memorizing regular and random patterns can be applied to memorize complex patterns, which are neither completely ordered nor disordered. The minimal perceptrons for describing complex patterns should require H between H_1 in Eq. (4) for regular patterns and H_2 in Eq. (6) for random patterns (Fig. 4):

$$H = (1 - \lambda)H_1 + \lambda H_2 \quad (7)$$

with $0 < \lambda < 1$. A complex pattern ξ can be decomposed into ordered elements, projected on a regular template pattern ξ_1 , and disordered elements, impurities deviated from the template. The Hamming distance between ξ and ξ_1 gives the number of impurities: $D = \sum_{\mu=1}^M (1 - \delta_{z^\mu, z_1^\mu})$. Here the impurity fraction is defined as $d \equiv D/M$. The dichotomy for the complex pattern uses H_1 hidden nodes for the regular elements and additional ΔH hidden nodes for the impurity elements. Then, the activation of the output node becomes $\sum_{j=1}^{H_1+\Delta H} v_j y_j - v_0$. Here the additional hidden nodes reverse the reference activation $\sum_{j=1}^{H_1} v_j y_j - v_0$ for the regular elements by adding a larger positive or negative value only for the impurity elements. The process of distinguishing dM impurity elements from $(1 - d)M$ regular ones is

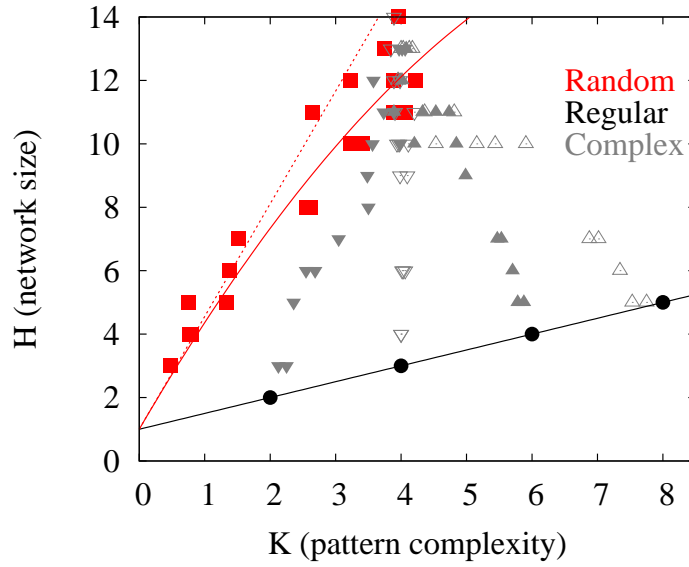


Figure 4: (Color online) Pattern complexity and network size. Random (red square), regular (black circle), and complex binary patterns (gray triangles) with an input size $N = 8$ are used. The complex patterns are generated by shuffling elements in the regular patterns of $(8, 6)$, $(8, 4)$, $(8, 2)$, and $(8, 0)$ pseudo-parity patterns (lower filled, lower empty, upper filled, and upper empty triangles, respectively). The lines are theoretical estimations for regular patterns [black, Eq. (4)] and for random patterns [dotted red, Eq 5, and solid red, Eq. (6)].

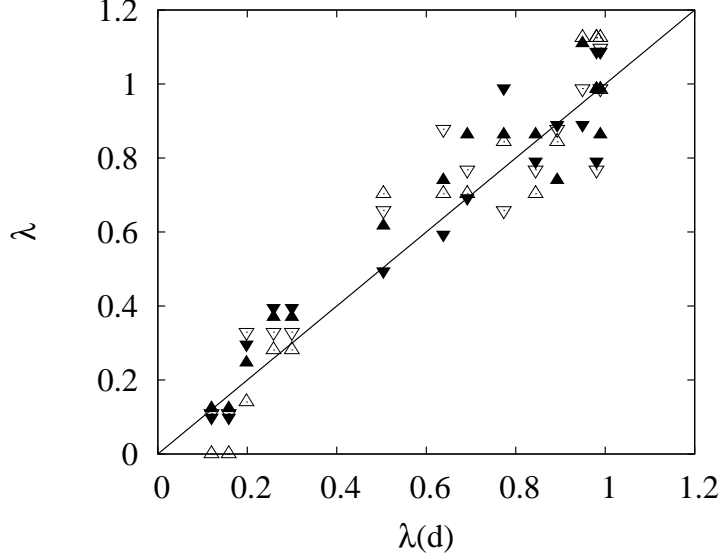


Figure 5: Interpolation of complex patterns. The interpolation parameter λ is compared with the predicted $\lambda(d)$ based on the Hamming distance d between complex patterns and regular patterns. The complex patterns are generated by shuffling elements in the regular patterns of $(8, 6)$, $(8, 4)$, $(8, 2)$, and $(8, 0)$ pseudo-parity patterns (lower filled, lower empty, upper filled, and upper empty triangles, respectively).

the same as the process of distinguishing pM black elements from $(1 - p)M$ white ones. Thus the formula, $H_2[K(p)]$ in Eq. (6), can be applied to estimate $\Delta H = H_2(d)$.

In the absence of impurity ($d = 0$), the pattern corresponds to a regular pattern ($H = H_1$). On the other hand, when the impurity is maximal ($d = p$), the required number of hidden nodes should become $H = H_2(p)$. This constraint introduces a scale factor $[H_2(p) - H_1]/H_2(p)$ for ΔH . Therefore, the minimum H for complex patterns is

$$H = H_1 + H_2(d) \cdot \frac{H_2(p) - H_1}{H_2(p)}. \quad (8)$$

The comparison of this equation with Eq. (7) gives the linear factor,

$$\lambda(d) = \frac{H_2(d)}{H_2(p)}. \quad (9)$$

Indeed, the Hamming distance d can be used to predict the linear factor λ (Fig. 5) and allows to estimate H in Eq. (7) for complex patterns.

6. Conclusion

Perceptrons encode a pattern into their weights and biases, which amounts to data compression in information theory [17]. In machine learning, determining the appropriate size of perceptrons to encode patterns is an old unsolved problem. Here we introduced a simple complexity index for patterns, and derived a minimum number of hidden nodes, dependent on this complexity index, for memorizing patterns. However, further study is needed to examine how additional nodes and complexities of network structure affect the robustness and adaptability of pattern storage.

7. Acknowledgements

We thank V. Periwal for discussions and hospitality during our stay at National Institutes of Health in the US. M. Pastor acknowledges the Office of Research, University of San Carlos, Cebu City, Philippines. This research was supported by Basic Science Research Program through the National Foundation of Korea funded by the Ministry of Science, ICT & Future Planning (No. 2013R1A1A1006655) and by the Max Planck Society, the Korea Ministry of Education, Science and Technology, Gyeongsangbuk-Do and Pohang City.

References

- [1] Ou G, Murphey YL. Multi-class pattern classification using neural networks. *Pattern Recognition* 2007;40:4–18.
- [2] Monterola C, Zapotocky M. Noise-enhanced categorization in recurrently reconnected neural network. *Physical Review E* 2005;71:036164. doi:10.1103/PhysRevE.71.036134.
- [3] Hopfield J. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of National Academy of Sciences* 1982;79:2554–8.
- [4] Abbott L, Arian Y. Storage capacity of generalized networks. *Physical Review A* 1987;36:5091. doi:10.1103/PhysRevA.36.5091.

- [5] Shiino M, Fukai T. Self-consistent signal-to-noise analysis of the statistical behavior of analog neural networks and enhancement of the storage capacity. *Physical Review E* 1993;48:867. doi:10.1103/PhysRevE.48.867.
- [6] Watkin T, Rau A. The statistical mechanics of learning a rule. *Reviews of Modern Physics* 1993;65:499–556. doi:10.1103/RevModPhys.65.499.
- [7] Blum A, Rivest R. Training a 3-node neural network is np-complete. *Neural Networks* 1992;5:117–27. doi:10.1016/S0893-6080(05)80010-3.
- [8] Murata N. YS, S.I. A. Network information criterion-determining the number of hidden units for an artificial neural network model. *IEEE Transactions on Neural Networks* 1994;5:865–72. doi:10.1109/72.329683.
- [9] Hirose Y. YK, S. H. Back-propagation algorithm which varies the number of hidden units. *Neural Networks* 1991;4:61–6. doi:10.1016/0893-6080(91)90032-Z.
- [10] Hunter D. Yu H. PMKJ, B. W. Selection of proper neural network sizes and architectures - a comparative study. *IEEE Transactions on Industrial Information* 2012;8:228–40. doi:10.1109/TII.2012.2187914.
- [11] Kwok T, Yeung D. Constructive algorithms for structure learning in feedforward neural networks for regression problems. *IEEE Transactions on Neural Networks* 1997;8:630–45. doi:10.1109/72.572102.
- [12] Reed R. Pruning algorithms-a survey. *IEEE Transactions on Neural Networks* 1993;4:740–7. doi:10.1109/72.248452.
- [13] Castellano G. FA, M. P. An iterative pruning algorithm for feedforward neural networks. *IEEE Transactions on Neural Networks* 1997;8:519–31. doi:10.1109/72.572092.
- [14] Shannon C. A mathematical theory of communication. *Bell System Technical Journal* 1948;27:379–423. doi:10.1002/j.1538-7305.1948.tb01338.x.

- [15] Silva Luiz Eduardo Virgilio ea. Multiscale entropy-based methods for heart rate variability complexity analysis. *Physica A: Statistical Mechanics and its Applications* 2015;422:143–52. doi:10.1016/j.physa.2014.12.011.
- [16] Larrondo H.A ea. Intensive statistical complexity measure of pseudo-random number generators. *Physica A: Statistical Mechanics and its Applications* 2005;356.1:133–8. doi:10.1016/j.physa.2005.05.025.
- [17] Cover T, Thomas J. *Elements of information theory*. Wiley-Interscience; 2006,.
- [18] LeCun Y. Bottou L. OG, K.R. M. *Efficient backdrop*. Springer Berlin Heidelberg; 1998,.
- [19] Haykin S. *Neural networks: a comprehensive foundation*. New York Prentice-Hall, Inc.; 1999,.
- [20] Cybenko G. Approximation by superpositions of a sigmoidal function. *Mathematics of Control Signals, and Systems* 1989;2:303–14. doi:10.1007/BF02551274.
- [21] Setiono R. On the solution of the parity problem by a single hidden layer feedforward neural network. *Neurocomputing* 1997;16:225–35. doi:10.1016/S0925-2312(97)00030-1.
- [22] Sontag E. Feedforward nets for interpolation and classification. *Journal of Computer and System Sciences* 1992;45:20–48. doi:10.1016/0022-0000(92)90039-L.
- [23] Rumelhart D. HG, R. W. Learning representations by back-propagating errors. *Nature* 1986;323:533–6.